4-1998

Item Estimates under Low-Stakes Conditions: How Should Omits Be Treated?

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Recommended Citation

DeMars, C. (1998, April). Item estimates under low-stakes conditions: How should omits be treated? Poster session presented at the annual meeting of the American Educational Research Association, San Diego

Item Estimates under Low-Stakes Conditions: How Should Omits Be Treated?

Christine DeMars
Michigan State University

Poster session presented at the annual meeting of the American Educational Research Association, San Diego, April 1998. (ERIC Document Reproduction Service No. ED 346 082)

ABSTRACT

Using data from a pilot test of science and math, item difficulties were estimated with a one-parameter model (partial-credit model for the multi-point items). Some items were multiple-choice items, and others were constructed-response items (open-ended). Four sets of estimates were obtained: estimates for males and females, and treating omitted items as incorrect and treating omitted items as not-presented (not-reached). Then, using data from an operational test (high-stakes, for diploma endorsement), the fit of these item estimates was assessed. In science, the fit was quite good under all conditions. In math, the fit was better for girls than for boys, the fit was better when omitted items were treated as not-presented, and the gender difference in fit was smaller when the omitted items were treated as not-presented.

The purpose of this study is to investigate the fit of response patterns observed under operational testing conditions to item estimates obtained under pilot conditions. The data were obtained on an untimed diploma-endorsement test with both multiple choice and constructed response items. Because the test did not have a time limit, omitted items could not be considered literally "not reached". On the operational test, items left unanswered are scored as wrong. However, non-response rates were higher on the pilot test, especially for the constructed response items. If omitted items on the pilot were scored as incorrect, and the non-response was due primarily to lack of motivation, not lack of knowledge, items with lower response rates would have higher difficulties (relative to other items) than they would under operational conditions. Groups of students with lower response rates would also have lower ability estimates. In this situation, it would be more accurate to treat omitted items as if they had not been administered (as is often done with strings of "not-reached" items at the end of low-stakes tests, especially where there is a fixed administration time). On the other hand, if non-response is due primarily to lack of knowledge (as it is assumed to be on the operational test), treating omitted items as incorrect will result in accurate estimates with smaller standard errors (because more responses are considered in the estimates).

Wolf, Smith, and Birnbaum (1995) found evidence of differential item functioning (DIF) between low-stakes and high-stakes groups on items which were classified as "mentally taxing". Low-stakes students did worse than expected (based on their overall abilities) on mentally taxing items. This would result in item misfit for these items if the item parameters were estimated in one group and tested for fit in the other (which is somewhat different from DIF techniques, but leads to similar conclusions). Though the items Wolf, Smith, and Birnbaum studied were all multiple choice items, constructed response items are generally more "taxing" (writing an answer requires more effort than choosing a pre-supplied response), so similar results may be found for constructed response items in this study.

Freund and Rock (1992) found that males appeared to be less motivated on the NAEP (as evidenced by marking distinct patterns on the answer sheet), a low-stakes test. If this low motivation is a larger factor on constructed response items (which take more effort), and if motivation increases when the stakes increase, performance on constructed response items may increase more than performance on multiple choice items for males, again leading to item misfit.

Test developers and researchers want estimated item difficulties from pilot tests to reflect the relative difficulties the items will have under operational conditions. If one type of item changes more than another, the test developers will not get the mix of item difficulties they intended. If the performance of one gender group changes more than another, estimates of group differences will not be accurate. Further, if there is an item by group by testing condition interaction, estimates of differential item functioning will not be the same.

Also, pilot tests share similarities with other low-stakes tests, such as the National Assessment of Educational Progress (NAEP) and various exams conducted by the International Education Association (IEA). On these examinations, students are informed that their scores will be anonymous and they receive no individual feedback. Achievement tests, for which students do receive individual scores but may perceive few consequences, and tests which have consequences at the school or district level, are between the extremes of high and low stakes for students. Findings about pilot tests could generalize to these other low-stakes situations; low-stakes tests may overestimate the difficulty of items, and the degree of the misestimate may depend on the format of the item and the gender group.

Method

Participants

Participants in the non-consequential (low stakes) test condition were students who participated in the pilot test (1994-95) of the science or math sections of the exam. For the forms studied, the students attended approximately 30 schools. Schools were randomly selected to participate, and students within schools were randomly assigned to test forms (two additional forms, to be used in other administrations, were administered to other students in these schools).

Participants in the consequential (high stakes) test condition were students who were tested during the spring 1996 (science) or 1997 (math) administration of the HSPT. Students who scored in the proficient category received state diploma endorsements. For this study, only the students in the schools which participated in the pilot testing were selected.

Instrument

At the time the data were collected, the Michigan High School Proficiency Test (HSPT) had four components: mathematics, science, reading, and writing. Beginning with the graduating class of 1997, students who scored in the proficient category received state diploma endorsements. Students initially took the tests in their junior year of high school, and students who did not pass had opportunities for retaking the exam. The tests are *not* designed as minimum competency exams; they are intended to reflect high school level skills.

The HSPT was first administered as an operational test (leading to diploma endorsements) in the spring of 1996. The science section administered at this testing period and the math section from the next regular (not re-take) administration were chosen for this study -- in 1996, many of the tested students had diploma endorsements in math from an earlier test, so it was not high-stakes.

The science section of the HSPT consists of 42 multiple choice items and eight constructed response items. Only 34 of the multiple choice items were analyzed here because eight of the items were different on the pilot test (for equating purposes). The maximum point value of the constructed response items varies across forms (a different form is developed for each administration), but on this test form, each constructed response item was worth two points. The test is administered in a single session and can be completed by most students within two hours, though students are given additional time if needed.

Two of the constructed response items ask students to read and critique a scientific investigation; these items were highly related, both conceptually and statistically. The correlation between these two items was .60, compared to correlations of .19-.33 for all other pairs of constructed response items. Therefore, these items were treated as a single item worth four points (nine categories when 1/2 points were considered). Effectively, then, there were seven constructed response items. Two other constructed response items based on the same text passage, as well as several clusters of multiple choice items (each cluster was written to relate to a common scenario or graphic) might also be logically viewed as more interdependent than items on the test as a whole. To check for this possibility, Yen's Q3 (Yen, 1993), was calculated for all item pairs, using operational data. For this statistic, an expected score is calculated for each student on each item, based on an item response function (Yen developed the statistic for the three-parameter model, but here the one-parameter model was used, as explained in the analysis section). The residual between each student's observed item score and his/her predicted item score is then found, and these residuals are correlated for pairs of items. Yen (1993) suggested a cutoff of .20 in deciding whether the assumption of local independence was violated to an extent which would make a practical difference. After combining the two investigation items, only two pairs of items met this criteria for the HSPT science test. One pair of multiple choice items had a residual correlation of .23; these items were treated as a set (scored 0-2) in item and ability estimation. Another pair, with one multiple choice and one constructed response item (adjacent and on the same theme), had a residual correlation of .21. These items were left separate because later analyses required separate scores on the multiple choice and constructed response sections. Two other pairs had residual correlations between .10 and .20. One was another mixed-format pair, which was left as two separate items, and the other was a multiple choice pair, which was combined into a single item (scored 0-2). All other correlations were less than .10.

The math section of the HSPT consisted of 40 multiple choice items and 6 constructed response items (worth 2-5 points each). Only 32 of the multiple choice items were analyzed here because 8 items were not consistent from pilot to operational test. As in science, the constructed response items involve written responses, which may include a drawing or figure and usually require words to be answered completely. In these items, students are expected to explain their solutions and reasoning processes.

There was no reason to expect to find sizable residual correlations on the math section, but for consistency with the science analyses the local item independence assumption was checked with Q3. Again, the operational data were used for the item calibration, and all items were estimated together. One pair of constructed response items was on the edge of Yen's (1993) suggested limit of .20, but I did not combine them because there were already so few constructed response items. There were four multiple choice items with intercorrelated residuals: five of the correlations were greater than .20, and the sixth was greater than .10. These items covered three different content areas, and only two were adjacent. These items were summed and treated as one item with five ordered score categories.

Results

To see how well the pilot parameter estimates fit the operational data, item parameters were estimated from the pilot data, using the one-parameter model (partial-credit model for the polytomous items). This is the model which is used on the operational test. Initially, omitted items were treated as incorrect, because students had adequate time to attempt all items if they wished. Then the items and abilities were recalculated treating omitted items as if they had not been administered. Parameters were estimated separately for each gender, under both these conditions. Based on these item estimates, the abilities of the operational students were re-estimated. The fit of the operational students' responses to the pilot item estimates was measured by the standardized appropriateness index of Levine, Drasgow, and Williams (1985), a composite index based on how probable the student's response to each polychotomous item is, given the student's estimated ability and the item's estimated parameters. This index has been shown to be roughly normally distributed in actual use (Levine, Drasgow, & Williams, 1985). Values less than zero indicate lower than expected fit, and values less than -2.58 would be rare, occurring for only 0.5% of the students in a sample where the model fit. Table 1 shows the mean standardized appropriateness fit, as well as the percent of students with indices less than -2.58, for each condition.

Table 1 - Appropriateness Fit Index

	Girls	Boys

Science	mean	% <-2.58	mean	% <-2.58
omitted items scored as incorrect	-0.28	1.77%	-0.11	0.43%
omitted items treated as not-administered	-0.25	1.43%	-0.05	0.36%
Math				
omitted items scored as incorrect	-0.43	8.06%	-0.53	9.07%
omitted items treated as not-administered	-0.24	7.09%	-0.25	7.26%

In science, the low-stakes item estimates fit better for the boys than for the girls. Fewer students were judged as misfitting when the item estimates were obtained by treating the omits as not-administered in the low-stakes test (though they were treated as incorrect on the high-stakes test). In math, the low-stakes item estimates fit better for girls than for boys, especially when omitted items were scored as incorrect. For both genders, item estimates fit better when omits were treated as not-administered. However, all sets had very large numbers of misfitting persons.

The appropriateness index was also calculated using the operational item estimates (ignoring gender), to serve as a baseline. The distributions were not precisely normal: In science, the standard deviation was 0.90, the skew was -0.50, and the kurtosis was 0.48. In math, the standard deviation was 0.99 (essentially 1), the skew was -0.61, and the kurtosis was 0.46. The mean was 0.01 in science and 0.04 in math, and only 0.7% of the students in science and 1.60% in math had values less than -2.58 (somewhat more than the 0.5% expected if normally distributed).

The appropriateness index targets students whose responses poorly fit the expected pattern. Another index, OUTFIT (Wright & Masters, 1982), targets items which fit the model poorly. The OUTFIT statistic was calculated based on the differences between observed and expected scores. For item i,

$$\text{OUTFIT}_i = \Sigma z^2_{ni}/N, \qquad (1)$$

where N is the total number of persons,

$$z_{ni} = \frac{x_{ni} - E_{ni}}{\sqrt{\sum_{k=0}^{m} (k - E_{ni})^2 \pi_{ni}}},$$

x_{ni} is the observed score for person n on item i,
k is the score point (range 0-m),
π_{nik} is the probability of person n scoring k on item i, and

E_{ni} (the expected score for person n on item i) = $\sum_{k=0}^{m} k\pi_{nik}$.

OUTFIT is more sensitive to unexpected outliers then a related measure, INFIT, in which each residual is weighted by its variance. Someone who has a high probability of a high score on an item but earns a low score instead (or someone who has a low probability of a high score but nevertheless obtains one) will have a larger impact on the OUTFIT statistic than on the INFIT statistic.

The four sets of item estimates (gender by method of treating omits) based on the pilot data and responses from the operational data were used to generate four OUTFIT indices for each item. The expected value of this index is one. Values greater than one indicate many responses are unexpected, and values less than one indicate responses are "too predictable"--there is not as much unexplained variance as expected. For every set, one science item (a constructed response item) had OUTFIT greater than 1.5 (50% greater than expected with good fit). OUTFIT for this item was larger for girls than boys, and it was larger when omits were treated as missing (not wrong) in the pilot data (recall that omits were always treated as wrong in the operational data). In math, two constructed response items had OUTFIT greater than 1.5 for all four sets, a third constructed response item had OUTFIT of 1.46 for one group and above 1.5 for the other three, and the multiple choice item composed of four individual items had OUTFITs between 1.28 and 1.52. Averages for each of the four sets, by item type, are displayed in Table 2.

Table 2 - Average OUTFIT

	Boys				Girls			
	omits as incorrect		omits as missing		omits as incorrect		omits as missing	
Science	mean	sd	mean	sd	mean	sd	mean	sd
Multiple Choice	0.9827	0.1894	0.9884	0.1972	1.0296	0.1306	1.0371	0.1391
Constructed Response	0.9998	0.2971	1.0631	0.2375	1.0407	0.2922	1.0984	0.2457
Math								
Multiple Choice	1.0445	0.1637	1.0051	0.1400	1.0646	0.1566	1.0387	0.1296
Constructed Response	1.6634	0.8989	1.4258	0.5221	1.3589	0.5250	1.2876	0.4455

In science, misfit was higher when omits were treated as not-administered (opposite the finding for the appropriateness index), and the gap was larger on constructed response items than on multiple choice items. OUTFIT was greater for girls than for boys.

In math, for both genders, there seemed to be an interaction between treatment of omits and response format; treating omits as not-administered improved fit more for the constructed response items than for the multiple choice items. This interaction was more extreme for boys than for girls.

Both the person-fit and the item-fit were extremely poor in math. One possible reason is because the equal-slopes model was more problematic in math than it was in science. Half-points were possible for the constructed-response items, and in order to utilize all information from these half-points and yet avoid overweighting the constructed-response items, the slope of each constructed-response item was set to half the slope of each multiple choice item for this study (the slope determines the weight of the item, and each category of the constructed response items was to be weighted a half point). To check the appropriateness of this ratio, a two-parameter model was run, using the high-stakes data. When the slopes were free to vary, the ratio of the average constructed response slope to the average multiple choice slope was about 0.6 in science, fairly close to the imposed 0.5 ratio. In math, however, this ratio was approximately 0.3. This probably contributed to the poor fit. The average OUTFIT for the constructed response items based on the operational item estimates was 1.265, which was better than the OUTFIT based on the pilot item estimates but still quite high for an average. However, the person-fit of the operational data using operational item estimates, as described above, was considerably better. Much of the poor fit, therefore, seems to be due to differences between pilot and operational responses.

Summary and Discussion

Because boys were less likely to respond to constructed response items, I expected that, for boys, ignoring omitted items on the pilot test when estimating item parameters would lead to item estimates which fit the operational data better than parameters estimated when omitted items were scored zero on the pilot test. In science, ignoring omitted items (treating them as if they had not been administered), improved person-fit (the appropriateness fit index) only slightly, and it led to somewhat poorer item-fit for the constructed-response items. These differences were small, and probably have little practical meaning. Fit (both item-fit and person-fit) was slightly worse for girls than boys, which was an unexpected finding. Because boys seemed to be less motivated on the pilot test (as evidenced by lower response rates on the constructed response items), I had expected them to have more idiosyncratic responses which would lead to poorer item estimates.

In math, my expectations held. Item fit and person fit both were better when omitted items were treated as not-administered, especially for boys and constructed response items. In math, the fit was generally much worse than it was in science, so there was more opportunity for differences. Also, in math non-response was higher, so method of treating omits would be expected to make more of a difference. The fit of the persons and of the constructed response items, though, regardless of how omits were treated, was remarkably poor. About 14 to 18 times as many people had very poor fit (z-scores less than -2.58), and this was mostly due to the poor fit of the constructed response items. The average outfit of the six constructed response items was 1.29 - 1.66 (depending on the group and conditions). For an individual item this would be 30% higher than expected if there were good fit, and for an average across multiple items it is quite high. This poor fit was partly due to the constraint of equal slopes--when the slopes were free to vary, the slopes of the constructed response items were not as steep as those of the multiple choice items.

One limitation to this study is that there was no measure of motivation other than tendency to omit constructed response items. As noted earlier, students may omit items for other reasons, such as little or no knowledge of the correct answer. It would be useful to have data on some other measure of motivation to separate these reasons for omission. Also, students likely varied in how important they felt the diploma endorsements were, depending on their interpretations of the messages they received from the schools, parents, other students, and the media about the diploma endorsements. These student beliefs, in turn, had an impact on how motivated the students were. If item estimates could be made separately for students who were unmotivated, it would reveal whether the poor fit in math was related to motivation or other factors.

References

Freund, D. S., & Rock, D. A. (1992, April). A preliminary investigation of pattern-marking in 1990 NAEP data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 347189)

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. Applied Measurement in Education, 8, 341-351.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

ABSTRACT

Using data from a pilot test of science and math, item difficulties were estimated with a one-parameter model (partial-credit model for the multi-point items). Some items were multiple-choice items, and others were constructed-response items (open-ended). Four sets of estimates were obtained: estimates for males and females, and treating omitted items as incorrect and treating omitted items as not-

presented (not-reached). Then, using data from an operational test (high-stakes, for diploma endorsement), the fit of these item estimates was assessed. In science, the fit was quite good under all conditions. In math, the fit was better for girls than for boys, the fit was better when omitted items were treated as not-presented, and the gender difference in fit was smaller when the omitted items were treated as not-presented.

Item estimates were based on pilot data.

Fit statistics were based on high-stakes data (for diploma endorsement.

OUTFIT is more influenced by extreme outliers than INFIT is (Wright & Masters).